arXiv:1004.4329v1 [cs.NA] 25 Apr 2010

The Journal of Fourier Analysis and Applications

Analysis of Basis Pursuit Via Capacity Sets

Joseph Shtok and Michael Elad

The Computer Science Department,

The Technion - Israel Institute of Technology, Haifa 32000 Israel,

email: [shtok,elad]@cs.technion.ac.il.

ABSTRACT. *Finding the sparsest solution α for an under-determined linear system of equations $\mathbf{D}\alpha = \mathbf{s}$ is of interest in many applications. This problem is known to be NP-hard. Recent work studied conditions on the support size of α that allow its recovery using ℓ_1-minimization, via the Basis Pursuit algorithm. These conditions are often relying on a scalar property of $\mathbf{D}$ called the mutual-coherence. In this work we introduce an alternative set of features of an arbitrarily given $\mathbf{D}$, called the* **capacity sets**. *We show how those could be used to analyze the performance of the basis pursuit, leading to improved bounds and predictions of performance. Both theoretical and numerical methods are presented, all using the capacity values, and shown to lead to improved assessments of the basis pursuit success in finding the sparest solution of $\mathbf{D}\alpha = \mathbf{s}$.*

1. Introduction

A powerful trend in signal processing that has evolved in recent years is the use of redundant dictionaries, rather than just bases, for a sparse representation of signals (images, sound tracks, and more). In such

Math Subject Classifications: 68P30, 68W25.
Keywords and Phrases: sparse representations, ℓ_1-reconstruction, Basis Pursuit, random support, capacity sets.

© 2004 Birkhäuser Boston. All rights reserved
ISSN 1069-5869 DOI: jfaatest- 05/28/04

a setting, we consider a linear equation $\mathbf{s} = \mathbf{D}\alpha$, where $\mathbf{s}$ is a given signal, $\mathbf{D}$ is the representation dictionary, and α is the signal's representation. The matrix $\mathbf{D}$ is a general full rank $N \times L$ matrix, where $L > N$, assumed to have ℓ_2 normalized columns. The number of non-zero elements in the coefficient vector α is measured by the ℓ_0-norm, $\|\cdot\|_0$, on $\mathbb{R}^L$. The goal is to find, within the $(L-N)$-dimensional affine space of the solutions for this equation, the sparsest representation for $\mathbf{s}$, i.e. one which has the least number of non-zero entries. This goal is formalized by the following optimization problem:

$$(P_0): \qquad \operatorname{Arg}\min_{\alpha \in \mathbb{R}^L} \|\alpha\|_0 \;\; s.t. \;\; \mathbf{D}\alpha = \mathbf{s}.$$

In this paper, we consider the signals for which the solution of (P_0) is unique, and we define $\mathcal{S}(\mathbf{D})$ as the family of such signals. We denote $\Omega = \{1, ..., L\}$, and refer to the support of the vector $\alpha = (\alpha_1, ..., \alpha_L)^T$ as the set $\Gamma = supp(\alpha) = \{n \in \Omega \mid \alpha_n \neq 0\}$.

The problem (P_0) is NP-hard, demanding an exhaustive search over all the subsets of columns of $\mathbf{D}$ [16]. One of the most effective techniques to approximate its solution is the convex relaxation of the ℓ_0-norm. It uses the ℓ_1-norm, the closest convex norm on $\mathbb{R}^L$:

$$(P_1): \qquad \operatorname{Arg}\min_{\alpha \in \mathbb{R}^L} \|\alpha\|_1 \;\; s.t. \;\; \mathbf{D}\alpha = \mathbf{s}.$$

The solution of (P_1) is carried out by linear programming. We are interested in signals $\mathbf{s} \in \mathcal{S}(\mathbf{D})$ for which the solutions of (P_0) and (P_1) coincide. The idea of using (P_1) to find the sparsest solution is called Basis Pursuit (BP), as coined by Chen, Donoho and Saunders [4, 5].

Let α be a representation of $\mathbf{s}$, with support $\Gamma = supp(\alpha) \subset \Omega$. The matrix $\mathbf{D}_\Gamma$ is a matrix of size $N \times |\Gamma|$ containing the columns (also

referred to as atoms) of $\mathbf{D}$ used for the construction of $\mathbf{s}$. This matrix is necessarily full-rank (with rank equals $|\Gamma|$). Knowing the support Γ suffices to enable perfect recovery of α, and thus our interest is confined to the ability to recover the support Γ.

Definition 1.1. *A subset $\Gamma \subset \Omega$ is called ℓ_1-**reconstructible** with respect to the dictionary $\mathbf{D}$ if the solution of (P_1) coincides with the solution of (P_0) for every signal $s \in \mathcal{S}(\mathbf{D})$ that admits a representation with the support Γ.*

The main task of the paper is to obtain conditions on support sizes which imply that they are ℓ_1-reconstructible. For any specific support $\Gamma \subset \Omega$ there exists a straightforward (yet exhaustive) test whether it admits recovery by BP – simply apply BP to the finite family of signals $\mathbf{s} = \mathbf{D}\alpha$ generated from coefficient vectors α with the support Γ covering all possible sign patterns (i.e. $2^{|\Gamma|}$ such tests[1]). If the recovery succeeds for all these choices of α, it will also succeed for any other representation with support Γ [9, 15].

Clearly, such a testing approach is impractical in most cases. If we aim to find the prospects of success of the BP for a fixed cardinality $|\Gamma|$, this requires a set of tests as described above per each possible support Γ having such a cardinality, and this implies a need for approximately $L^{|\Gamma|}$ groups of tests. Thus, the exhaustive approach should be replaced either by a random set of tests with empirical claims, or a theoretical study.

Within the theoretical attempts to estimate the power of the BP, two approaches are distinguished in the existing literature. Earlier

[1] In fact, half of this amount is required because if α is reconstructible, then so is $-\alpha$.

work carried out the worst case analysis for a given dictionary, providing conditions on the support cardinality that guarantee that any support satisfying them is ℓ_1-reconstructible [8, 9, 11, 12, 13, 20]. These conditions are often very restrictive and far from empirical evidence. Another, more recent, approach presents a probabilistic analysis, providing conditions for special families of dictionaries under which *most* signals of a given cardinality are ℓ_1-reconstructible [1, 2, 6, 10, 19]. The results depict a general asymptotic behavior with regard to the sparse support recovery.

In both worst-case and probabilistic-analysis branches of work, many classical results rely heavily on a scalar feature of the dictionary, known as the *mutual-coherence* [8, 12, 13, 20]. A related measure also used is the Babel function [8, 20]. More recent work employs the Restricted Isometry Property (RIP) [3]. The information carried by all these measures is very pessimistic; furthermore, the RIP is very expensive computationally and mainly used for theoretical analysis. In this work we set to improve the existing worst case results for a given general dictionary $\mathbf{D}$, as reported in [8, 12, 13, 20]. We achieve this progress by replacing the above-mentioned with a set of alternative features that we refer to as the *capacity sets* of the dictionary. A thorough computational analysis of $\mathbf{D}$ and probabilistic tools are applied to the problem, leading to improved probabilistic bounds.

In the next section we recall the existing theoretical results concerning ℓ_1-recovery as a function of the support cardinality. In section 3 we define two versions of the *capacity set* and present the main theoretical results of this paper using these features. Section 4 expands on the above results by providing two numerical algorithms using the

capacity sets. Section 5 provides an overall comparison of the various methods presented in this work to assess the performance of BP for several test-cases.

2. Background

Most known results on sparsity rely on the *mutual-coherence*, denoted as μ, of the dictionary. This is the maximum of the inner products between the columns: $\mu = \max_{i \neq j \in \Omega} | < \mathbf{d}_i, \mathbf{d}_j > |$. This correlation between the columns, reflected in its worst value by μ, helps establishing the "safe zone" for the support sizes, where both the uniqueness of sparsest representation and its ℓ_1-recovery can be guaranteed.

For $\mathbf{D} = [\mathbf{\Phi}_1, \mathbf{\Phi}_2]$ a pair of orthonormal bases, the following sufficient condition for Γ to be ℓ_1-reconstructible is proven in [11]:

$$|\Gamma| \leq \frac{\sqrt{2} - 0.5}{\mu} \quad .$$

Donoho and Elad in [8] treat a general dictionary $\mathbf{D}$. They define the problem

$$(C_\Gamma): \qquad \max_{\delta \in Null(\mathbf{D})} \sum_{k \in \Gamma} |\delta_k| \quad s.t. \quad \|\delta\|_1 = 1 \ , \tag{2.1}$$

and show that its solution is intimately tied to the ability to recover the support Γ, by the following lemma:

Lemma 2.1. *([8], Lemma 2) A sufficient condition on the support Γ to be ℓ_1-reconstructible is*

$$val(C_\Gamma) < \frac{1}{2}. \tag{2.2}$$

This criteria is used to prove the following theorem:

Theorem 2.2. *([8], Theorem 7) A sufficient condition on a support $\Gamma \subset \Omega$ to be ℓ_1-reconstructible is*

$$|\Gamma| < \frac{1}{2}\left(1 + \frac{1}{\mu}\right). \tag{2.3}$$

Typically, the coherence behaves at best like $\mathcal{O}(\frac{1}{\sqrt{N}})$, hence the results stated above predict quite weak ℓ_1-recovery, which is refuted by the empirical evidence: usually BP recovers supports of size proportional to N (and not its squared-root).

A generalization of the coherence is introduced in [8] and later used by J. Tropp in [20]: for any $0 \leq m \leq L$, the Babel function $\mu_1(m)$ is defined by

$$\mu_1(m) = \max_{|\Lambda|=m} \max_{\eta \in \Omega \backslash \Lambda} \sum_{\lambda \in \Lambda} | < \phi_\lambda, \phi_\eta > |.$$

In terms of this function, a support of size m is proven to be ℓ_1-reconstructible provided the following inequality holds [20]:

$$\mu_1(m-1) + \mu_1(m) < 1.$$

Unfortunately, in cases where the coherence μ is close to 1 (implying an existence of at least one problematic pair of atoms), the growth of $\mu_1(m)$ is too fast to provide any improvement.

Average case analysis improves the asymptotic bounds on reconstructible support sizes. The work in [2] shows that for the dictionary $\mathbf{D} = [\mathbf{I}, \mathbf{F}^*]$, where $\mathbf{F}$ is the Fourier transform, random uniformly sampled support admits ℓ_1-recovery with high probability if (the expectation of) its cardinality is $\mathcal{O}(N/\log N)$, which improves the $\mathcal{O}(\sqrt{N})$ estimation of the worst case approach. For a general orthonormal pair, it is shown in ([2], Theorem 5.3) that most random supports which

cardinality behaving like $\mathcal{O}(1/(\mu^2 \log^6 N))$ admit recovery by BP. The $\log N$ appearing in these expressions is suspected by the authors of [2] to be unnecessary, which in effect turns this expression into $\mathcal{O}(N)$ (for incoherent dictionaries). A similar and related result, exhibiting the square of the mutual coherence in the denominator of the bound, appears in [19]. As such, this result is effective in cases where the dictionary is "uniformly coherent", and the methods employed are not very suitable for dictionaries with high coherence.

The idea that representations with cardinalities $\mathcal{O}(N)$ are ℓ_1 -reconstructible is supported by the results reported in [6, 7, 10]. This result is obtained for asymptotically growing dictionaries of size $N \times \delta N$ constructed by concatenating random vectors of unit l_2-norm, independently drawn from the uniform distribution. It is shown that all supports of size up to $\rho(\delta)N$ are ℓ_1-reconstructible with probability approaching 1. The work in [7, 10] provides theoretical assessments for $\rho(\delta)$, based on connection to study on neighborly polytopes. Despite being asymptotical, these results illuminate the empirically-supported evidence regarding the reconstruction abilities of minimal L_0-norm supports by linear programming.

As good as these results sound, they do not provide useful numerical information about the ability of ℓ_1-reconstruction applied to a specifically given dictionary $\mathbf{D}$ of certain size, which is a practical and central question in the application of BP. Such information can only be obtained today by results involving the coherence μ or its descendants. Thus, the gap is especially big when the dictionary is not uniformly coherent and when $\mu \gg \dfrac{1}{\sqrt{N}}$.

In this work we introduce new features of the dictionary $\mathbf{D}$, the

capacity sets. These features are obtained as the solutions to specific linear programming problems that probe the dictionary $\mathbf{D}$. We consider two such options: a vector of capacities $\mathbf{q}$ and a matrix $\mathbf{Q}$, as we shall explain in details in the next section. These features are used to develop novel analysis of BP performance as a function of the support's cardinality.

One interesting benefit of the proposed analysis is a better treatment of dictionaries which are not "uniformly coherent". In cases where there exists a small set of columns in $\mathbf{D}$ with strong linear dependency, the coherence and the babel function behave badly, tending to lead to overly pessimistic bounds. As we show, the use of the capacities leads in these cases to much better results. Besides that, the capacities are shown to be more delicate indicators of the dictionary, as reflected in a better prediction of the BP performance.

Use of *capacity sets* bridges the gap between purely theoretical estimations of the reconstructible support sizes for given dictionary $\mathbf{D}$, which are usually fast but provide pessimistic lower bound, and the empirical tests of $\mathbf{D}$, which give very accurate account on BP-reconstruction abilities, but are computationally prohibitive. We propose theoretical results and algorithms that employ the *capacity sets* to perform computational assessment of these abilities, which is fast relative to full empirical test and more optimistic than known practical formulae. The question of computational complexity is discussed in details in section 5.4.

3. Capacity Sets and Their Use

In this section we define two versions of the *capacity sets*, and state the main theoretical results that employ them for the analysis of the BP.

3.1 The Capacity Vector $\mathbf{q}$

The capacity vector consists of elements related to an intermediate tool used in the proof of Theorem 2.2 in [8]:

Definition 3.1. *The capacity vector* $\mathbf{q} = (q_1, ..., q_L)^T$ *of a dictionary* $\mathbf{D} \in \mathbb{R}^{N \times L}$ *is defined for all* $k \in \Omega$ *by*

$$q_k = \max_{\delta \in Null(D)} \delta_k \quad s.t. \quad \|\delta\|_1 = 1. \tag{3.1}$$

Computing the elements of $\mathbf{q}$ is relatively easy, and amounts to a simple set of L independent linear programming problems of the form

$$\hat{\mathbf{x}}_k = \mathrm{Arg}\min_{\mathbf{x}} \ ||\mathbf{x}||_1 \quad \text{subject to} \quad \mathbf{D}\mathbf{x} = \mathbf{0} \ \text{ and } \ x_k = 1,$$

and then assigning $q_k = 1/||\hat{\mathbf{x}}_k||_1$.

To see the equivalence of the two problems, notice that the vector $\tilde{\mathbf{x}}_k = \hat{\mathbf{x}}_k / \|\hat{\mathbf{x}}_k\|_1$ is an element of null space of $\mathbf{D}$ with unit ℓ_1-norm. Since $(\hat{\mathbf{x}}_k)_k = 1$ and $\|\hat{\mathbf{x}}_k\|_1$ is smallest possible, the value $q_k = 1/||\hat{\mathbf{x}}_k||_1 = (\tilde{\mathbf{x}}_k)_k$ is just the solution of 3.1.

Via Lemma 2.1, the definition of $\mathbf{q}$ provides a sufficient condition $\sum_{k \in \Gamma} q_k < \frac{1}{2}$ on a given support Γ to ensure its recovery by ℓ_1-minimization. Furthermore, by gathering the $|\Gamma|$ largest entries from $\mathbf{q}$, a simple generalization of Theorem 2.2 can be proposed. However, in this work we seek a better bound that takes into account the variety of possible supports, rather than the worst one. One such numerical technique is suggested in section 4, proposing a special quantization of

the values in $\mathbf{q}$ to obtain a lower bound on the fraction of support sizes which admit recovery by BP.

In this section we aim to obtain a more theoretically flavored result that uses $\mathbf{q}$. Denote by E_q the mean value of the capacity vector $\mathbf{q}$, and by σ_q^2 its variance $\frac{1}{L}\sum_{k\in\Omega}(q_k - E_q)^2$. The following theorem uses these quantities to evaluate the probability of ℓ_1-reconstruction for a given support size:

Theorem A. *For any $1 \leq \ell < \frac{1}{2E_q}$, a support Γ of size ℓ, sampled uniformly at random from Ω, admits ℓ_1-recovery with probability*

$$P(\ell) > \frac{\left(\frac{1}{2} - \ell E_q\right)^2}{\ell \sigma_q^2 + \left(\frac{1}{2} - \ell E_q\right)^2} \ . \tag{3.2}$$

In the special case of a constant capacity vector, the theorem boils down to support size threshold of $\frac{1}{2E_q}$, since then the variance becomes zero. We show in Section 3.2 that weakened version of Theorem A yields the classical threshold of $|\Gamma| < \frac{1}{2}\left(1 + \frac{1}{\mu}\right)$ (see Theorem 2.2).

Proof: We fix ℓ and chose subsets $\Lambda, \Gamma \subset \Omega$ according to two different probability models. The elements of Γ are chosen uniformly from Ω without replacement and form a set of ℓ distinct column indices. The ℓ elements of Λ are chosen uniformly with replacement (i.e. Λ is a multiset of size ℓ with possible duplicates). Now, define random variables

$$x_\ell = \sum_{k\in\Gamma} q_k, \quad y_\ell = \sum_{m\in\Lambda} q_m. \tag{3.3}$$

In these terms, the probability $P(\ell)$, defined in the statement of the theorem, is bounded below by $P(x_\ell < \frac{1}{2})$. In turn, we shall bound the probability $P(x_\ell < \frac{1}{2})$ by means of the Tchebychev inequality, which involves the mean and the variance of x_ℓ. These parameters

are easily computable for y_ℓ: by its definition, we have $\mathbb{E}(y_\ell) = \ell E_q$, $var(y_\ell) = \ell\sigma_q^2$. Our result is based on the following connection between the variables x_ℓ and y_ℓ, as shown in Appendix A:

$$\mathbb{E}(x_\ell) = \mathbb{E}(y_\ell) \;\text{ and }\; var(x_\ell) \le var(y_\ell). \tag{3.4}$$

Given any real scalar $a > 0$, the one-tailed version of the Tchebychev inequality [14] for x_ℓ reads

$$P\left(x_\ell - E_x \ge a\sigma_x\right) = P\left(x_\ell \ge E_x + a\sigma_x\right) \le \frac{1}{1+a^2}\ ,$$

where $E_x = \mathbb{E}(x_\ell)$, $\sigma_x^2 = var(x_\ell)$.

By (3.4), we substitute $E_x = \ell E_q$. Also, since a larger variance implies a lower probability, we put $\sqrt{\ell}\sigma_q$ instead of σ_x and obtain

$$P\left(x_\ell \ge \ell E_q + a\sqrt{\ell}\sigma_q\right) \le P\left(x_\ell \ge E_x + a\sigma_x\right) \le \frac{1}{1+a^2}.$$

The parameter a is chosen such that $\ell E_q + a\sqrt{\ell}\sigma_q = \frac{1}{2}$, leading to $a = (\frac{1}{2} - \ell E_q)/(\sqrt{\ell}\sigma_q)$. Note that the condition $a > 0$ translates to the requirement $\ell < \frac{1}{2E_q}$ as claimed in the theorem. In case it holds, we have

$$P\left(x_\ell \ge \frac{1}{2}\right) \le \frac{1}{1 + \frac{\left(\frac{1}{2} - \ell E_q\right)^2}{\ell\sigma_q^2}}\ ,$$

or put differently,

$$P(x_\ell < \frac{1}{2}) > 1 - \frac{1}{1 + \frac{\left(\frac{1}{2} - \ell E_q\right)^2}{\ell\sigma_q^2}} = \frac{\left(\frac{1}{2} - \ell E_q\right)^2}{\ell\sigma_q^2 + \left(\frac{1}{2} - \ell E_q\right)^2}\ ,$$

as stated by the theorem. □

3.2 From Capacity Vector to Coherence

We mentioned earlier that previous work often uses the *mutual coherence* to derive performance bounds on ℓ_1-reconstructible supports. The

relation between the capacities in $\mathbf{q}$ and the inner products between the dictionary atoms, $| < \mathbf{d}_i, \mathbf{d}_j > |$ has been already discussed in [8]. Given a dictionary $\mathbf{D}$, construct its Gram matrix as $\mathbf{G} = \mathbf{D}^T\mathbf{D}$. Define the sequence

$$\mu_k = \max_{i \neq k} \; |G_{i,k}| \;\; \text{for} \;\; k \in \Omega. \tag{3.5}$$

Namely, μ_k is the maximal value on the k-th column of $|\mathbf{G}|$, disregarding the main diagonal entry. As [8] shows, this sequence of values satisfies

$$q_k \leq \frac{\mu_k}{\mu_k + 1}.$$

Thus the condition $\sum_{k\in\Gamma} q_k < \frac{1}{2}$ can be replaced with $\sum_{k\in\Gamma} \frac{\mu_k}{\mu_k+1} < \frac{1}{2}$, leading of-course, to weaker bounds. Further relaxation

$$q_k \leq \frac{\mu_k}{\mu_k + 1} < \frac{\mu}{\mu + 1} \tag{3.6}$$

yields a constant capacity vector with entries of size $\frac{\mu}{\mu+1}$. Applying Theorem A to this vector we obtain, as a special case, the classical Theorem 2.2.

3.3 Using the Capacity Matrix Q

One problem with the capacity vector $\mathbf{q}$ is the independence with which its entries q_k are computed. This implies that one (or more) of the entries in $\mathbf{q}$ may become unnecessarily large, compared to the values obtained in Equation (2.1), causing a weaker bound. By working with pairs of such entries, one could in principle improve the obtained bounds. This leads us to the following definition:

Definition 3.2. *Denote by* Ω_2 *the set of indices* $\Omega_2 = \{(i,j)|\; i,j \in \Omega, i < j\}$. *The upper triangular capacity matrix* $\mathbf{Q} = \{Q_{i,j}\}$ *is the*

matrix with non-zero elements indexed by $(i, j) \in \Omega_2$*, defined as follows:*

$$Q_{i,j} = \max_{\delta \in Null(\mathbf{D})} \{\max(\delta_i + \delta_j, \delta_i - \delta_j)\} \ s.t. \ \ \|\delta\|_1 = 1.$$

Each of these entries can be computed by two independent linear programming problems of the form

$$\left\{ \begin{array}{l} \mathbf{x}^+_{(i,j)} = \operatorname{Arg\,min}_{\mathbf{x}} \ ||\mathbf{x}||_1 \ \text{ subject to } \ \mathbf{D}\mathbf{x} = \mathbf{0} \text{ and } x_i + x_j = 1 \\ \mathbf{x}^-_{(i,j)} = \operatorname{Arg\,min}_{\mathbf{x}} \ ||\mathbf{x}||_1 \ \text{ subject to } \ \mathbf{D}\mathbf{x} = \mathbf{0} \text{ and } x_i - x_j = 1 \end{array} \right\}$$

and then assigning $Q_{i,j} = 1/\min(||\hat{\mathbf{x}}^+_{(i,j)}||_1, ||\hat{\mathbf{x}}^-_{(i,j)}||_1)$.

As in section 3.1, the obtained values $Q_{i,j}$ could be used to form an improved worst-case bound for Lemma 2.1 and consequently for Theorem 2.2: Let $\Gamma \subset \Omega$ be a randomly chosen support of size[2] $\ell = 2n$. By definition, the non-zero elements of $\mathbf{Q}$ satisfy

$$\max_{\substack{\delta \in Null(D) \\ \|\delta\|_1 = 1}} |\delta_i| + |\delta_j| = Q_{i,j} \leq \max_{\substack{\delta \in Null(D) \\ \|\delta\|_1 = 1}} |\delta_i| + \max_{\substack{\delta \in Null(D) \\ \|\delta\|_1 = 1}} |\delta_j| = q_i + q_j.$$

Thus the values $Q_{i,j}$ can be used in the evaluation of an upper bound on C_Γ. To any partition $\mathcal{I}$ of Γ into disjoint pairs there corresponds the sum $\sum_{(k_1,k_2) \in \mathcal{I}} Q_{k_1,k_2}$ that bounds the value of C_Γ from above. Therefore, Γ is ℓ_1-reconstructible if there exists such a partition satisfying $\sum_{(k_1,k_2) \in \mathcal{I}} Q_{k_1,k_2} < \frac{1}{2}$. Naturally, among all such possible partitions, we are interested in the one that leads to the smallest sum.

Just one glance at the values of $\mathbf{Q}$ gives a lower bound for sizes of ℓ_1-reconstructible subsets: namely, if $\max(\mathbf{Q}) \leq \frac{1}{\ell}$, then a sum of any $\ell/2$ of its elements does not exceed $1/2$; hence any subset of columns of

[2] We consider hereafter even support sizes. Generalization to odd ones is relatively simple, requiring use of one entry from $\mathbf{q}$. We omit this discussion for simplicity.

size up to ℓ is guaranteed to be recovered by BP. Conjecture B below estimates the uncertainty caused by replacing $\max(\mathbf{Q})$ with $mean(\mathbf{Q})$. Some numerical techniques based on $\mathbf{Q}$ are described in section 4.

Here we concentrate again on a theoretical bound that uses $\mathbf{Q}$, similar to the one proposed in Theorem A with few necessary modifications.

We arrange the values $\{Q_{i,j} \mid i < j \in \Omega\}$ of the Capacity matrix in a vector $\mathbf{Q}^V$. Denote by E_Q the mean value of $\mathbf{Q}^V$, and by σ_Q^2 its variance, $\sigma_Q^2 = \frac{2}{L(L-1)} \sum_{i<j\in\Omega} (Q_{i,j} - E_Q)^2$. The following statement based on $\mathbf{Q}$ is similar to the one in Theorem A:

Conjecture B. [3] *For any $1 \leq \ell < \frac{1}{E_Q}$, a support Γ of even size ℓ, sampled uniformly at random from Ω, admits ℓ_1-recovery with probability*

$$P(\ell) > \frac{\left(\frac{1}{2} - \frac{\ell}{2}E_Q\right)^2}{\frac{\ell}{2}\sigma_Q^2 + \left(\frac{1}{2} - \frac{\ell}{2}E_Q\right)^2}. \tag{3.7}$$

Notice that the expression obtained in Equation(3.7) is the same as the one in (3.2), with ℓ replaced by $\ell/2$. Since E_Q and σ_Q refer to pairs, if $E_Q = 2E_q$ and $\sigma_Q^2 = 2\sigma_q^2$ the two bounds are the same. However, as we shall demonstrate in section 5, $E_Q < 2E_q$ and $\sigma_Q^2 < 2\sigma_q^2$ for random dictionaries, implying that this bound is indeed stronger.

Proof: Fix an even support size ℓ. In order to translate the condition $\sum_{(i,j)\in\mathcal{I}} Q_{i,j} < \frac{1}{2}$ to a probabilistic one, we use again the model involving a subset $\Gamma \subset \Omega$ of size ℓ which elements are chosen uniformly from Ω without replacement. Also, we let $\mathcal{I}$ be a random partition of

[3]This claim is a conjecture since it relies on a property that is used here without a proof. More on this is given in Appendix B.

the index set Γ into pairs. Based on these notions, we define a random variable $x_\ell = \sum_{(k_1,k_2)\in\mathcal{I}} Q_{k_1,k_2}$. In effect, x_ℓ is a sum of elements of $\mathbf{Q}$ randomly chosen "without replacement" in a stronger sense, i.e. not only the elements are not repeated, but two elements with common index are not allowed. The probability $P(\ell)$, defined in the statement of the theorem, is bounded below by $P(x_\ell < \frac{1}{2})$. This bound is not tight, since the support Γ is reconstructible if there exists *some* partition $\mathcal{I}^{opt}$ such that $\sum_{(k_1,k_2)\in\mathcal{I}^{opt}} Q_{k_1,k_2}$ drops below the half, while $P(x_\ell < \frac{1}{2})$ is only the probability this will happen for a *random* partition $\mathcal{I}$.

In order to analyze the variable x_ℓ we consider a multiset Φ of size $\frac{\ell}{2}$ chosen uniformly with replacement from $\mathbf{Q}^V$, and define the random variable y_ℓ to be its sum, $y_\ell = \sum \Phi$. Then we have $\mathbb{E}(y_\ell) = \frac{\ell}{2}E_Q$, $var(y_\ell) = \frac{\ell}{2}\sigma_Q^2$.

The expectation of x_ℓ equals to that of y_ℓ, which is proven in Appendix B. Regarding the variance, we are making an assumption similar to 3.4:

$$var(x_\ell) \leq var(y_\ell). \tag{3.8}$$

We do not provide its proof and leave it as an open question at this stage. Empirical verification of this inequality is demonstrated in Appendix B.

Following the steps of Theorem A, given any real $a > 0$, the one-tailed version of the Tchebychev inequality [14] for x_ℓ reads

$$P\left(x_\ell \geq \frac{\ell}{2}E_Q + a\sqrt{\frac{\ell}{2}}\sigma_Q\right) \leq \frac{1}{1+a^2}.$$

The parameter a is chosen such that $\frac{\ell}{2}E_Q + a\sqrt{\frac{\ell}{2}}\sigma_Q = \frac{1}{2}$, leading to $a = (\frac{1}{2} - \frac{\ell}{2}E_Q)/(\sqrt{\frac{\ell}{2}}\sigma_Q)$, implying that we should require $\ell < \frac{1}{E_Q}$ to

get $a > 0$. This leads to

$$P\left(x_\ell \geq \frac{1}{2}\right) \leq \frac{1}{1+\frac{\left(\frac{1}{2}-\frac{\ell}{2}E_Q\right)^2}{\frac{\ell}{2}\sigma_Q^2}} \quad ,$$

or put differently,

$$P(x_\ell < \frac{1}{2}) > 1 - \frac{1}{1+\frac{\left(\frac{1}{2}-\frac{\ell}{2}E_Q\right)^2}{\frac{\ell}{2}\sigma_Q^2}} = \frac{\left(\frac{1}{2}-\frac{\ell}{2}E_Q\right)^2}{\frac{\ell}{2}\sigma_Q^2+\left(\frac{1}{2}-\frac{\ell}{2}E_Q\right)^2} \quad ,$$

as stated in the theorem. □

4. Numerical Algorithms

Given the capacity vector $\mathbf{q}$ (or its weaker version as described in section 3.2) or matrix $\mathbf{Q}$, we can use Theorems A and B to predict the ℓ_1-reconstructible supports, and show lower bounds of the probability for success as a function of the support size ℓ. However, we can alternatively evaluate these probabilities numerically, provided that there are shortcuts that avoid the exponential growth in support possibilities. This leads us to the following two algorithms.

4.1 A Fast Combinatorial Count Using q

Below we propose an algorithm which provides worst-case bounds on reconstructible support sizes. We would like to establish the fraction of the total number of supports Γ of size ℓ that satisfy $val(C_\Gamma) < \frac{1}{2}$. Testing the sufficient condition $\sum_{k\in\Gamma} q_k < \frac{1}{2}$ for every single Γ requires $\mathcal{O}(L^\ell)$ flops, which is prohibitive. Instead, we propose to perform a quantization of the entries of $\mathbf{q}$ to d distinct values, and lead to a more reasonable computational process.

Suppose we are given a partition $\Lambda = \{\Lambda_i\}_{i=1}^d$ of Ω into d disjoint clusters, such that $\Omega = \bigcup_{i=1}^d \Lambda_i$. The corresponding quantized values

in $\mathbf{q}$ are denoted by $\{q_\Lambda^i\}$, each set to be the maximal in its subset, $\{q_\Lambda^i = \max_{k\in\Lambda_i}(q_k) \mid 1 \le i \le d\}$.

Given the quantization parameters $\Lambda = \{\Lambda_i,\ q_\Lambda^i\}_{i=1}^d$, every ℓ-sized support $\Gamma \in \Omega$ can be described as the union $\bigcup_{i=1}^d \Gamma_i$, where $\Gamma_i \subseteq \Lambda_i$ is the subset of indices in Γ allocated to the quantized value q_Λ^i. Thus, the sum $\sum_{k\in\Gamma} q_i$ can be replaced by a larger sum, $\sum_{i=1}^d |\Gamma_i| q_\Lambda^i$.

In order to test all possible supports $\Gamma \in \Omega$ of size ℓ, a combinatorial count of all sequences $p = (p_1, ..., p_d)$ is performed, such that $0 \le |p_i| \le |\Lambda_i|$ and $\sum_{i=1}^d |p_i| = \ell$. For each of these we evaluate $\sum_{i=1}^d |p_i| q_\Lambda^i$ and count the relative number of those[4] below $\frac{1}{2}$. The complexity of such computation does not exceed $\mathcal{O}\left((\frac{L}{d})^d\right)$.

As to the choice of the quantization parameters $\Lambda = \{\Lambda_i,\ q_\Lambda^i\}_{i=1}^d$, as said above, we let $q_\Lambda^i = \max_{k\in\Lambda_i} q_k$ to guarantee that the evaluated summations are considering a worst-case scenario. The clustering is done by an attempt to minimize the function

$$f\left(\{\Lambda_i,\ q_\Lambda^i\}_{i=1}^d\right) = \sum_{i=1}^d \left(|\Lambda_i| q_\Lambda^i - \sum_{k\in\Lambda_i} q_k \right). \tag{4.1}$$

The difference $|\Lambda_i| q_\Lambda^i - \sum_{k\in\Lambda_i} q_k$ is the quantization error for the elements in the subset Λ_i, and the above error simply sums these values.

The minimization of $f\left(\{\Lambda_i,\ q_\Lambda^i\}_{i=1}^d\right)$ can be done exhaustively in case d is small – in our experiments we have used $d = 3$ implying that the above requires $\mathcal{O}(L^3)$ flops. For larger values of d a sequential algorithm that chooses Λ_i can be proposed, separating the set Ω to two parts, and proceeding in a tree and greedy separation scheme.

Computationally, the results of the combinatorial count are very close to those predicted by Theorem A. Therefore, this method serves as

[4]Each instance must be weighted by the number of its possible occurrences.

a supporting evidence for the probabilistic approach taken in Theorem A, but its numerical output is omitted from our display of experimental results in section 5.

4.2 A Sampling Algorithm Using Q

An alternative to Conjecture B is a direct evaluation of ℓ_1-reconstructible supports Γ of cardinality ℓ, by the following stages:

- We draw $M \gg L$ such supports $\{\Gamma_i\}_{i=1}^{M}$.
- For each Γ_i we seek to find a partition $\mathcal{I}_i$ that leads to the smallest value of $\sum_{(k,l)\in\mathcal{I}} Q_{k,l}$. While finding the best such partition is combinatorial in complexity, we use an approximate greedy algorithm of complexity $\mathcal{O}(\ell^2 \cdot log(\ell))$ which computes the following suboptimal partition:

  1. Begin with empty set $\mathcal{I}$ of pairs.

  2. denote by $\mathbf{Q}_{res}$ the sub-matrix of $\mathbf{Q}$ which rows an columns consist of only those indices from $|\Gamma|$ which do not occur in $\mathcal{I}$. Retrieve the couple $(i_0, j_0), (i_1, j_1)$ of index pairs which minimize the sum $\mathbf{Q}(i_0, j_0) + \mathbf{Q}(i_1, j_1)$ over $\mathbf{Q}_{res}$.

  3. joint the couple $(i_0, j_0), (i_1, j_1)$ to $\mathcal{I}$ and return to item 2 while $\mathbf{Q}_{res}$ is nonempty.

  Therefore, the algorithm is, in a sense, "second-order greedy", i.e. at each step the least-sum couple of values from $\mathbf{Q}$, rather than least single value, is extracted. Possibly, better algorithms will improve the performance of this scheme, but we believe it to be quite close to optimal, while keeping low computational costs. The fact such partition can be found in $\mathcal{O}(\ell^2 \cdot log(\ell))$

follows from the next combinatorial claim: let (i^*, j^*) be the index pair of minimal value in submatrix of $\mathbf{Q}$ supported on $|\Gamma|$. Then both i^*, j^* necessarily present among indices (i_0, j_0, i_1, j_1) defined above.

- Given the partition $\mathcal{I}$, test $\sum_{(k,l)\in\mathcal{I}} Q_{k,l} < \frac{1}{2}$. Accumulate the relative number of such occurrences over the collection $\{\Gamma_i\}_{i=1}^{M}$.

The fact that this method relies on capacity values implies that the predicted performance is expected to be weaker compared to the true behavior of BP. Nevertheless, among the various methods discussed thus far, this method is expected to be the most optimistic because it uses $\mathbf{Q}$ and not $\mathbf{q}$, and also because it does not build the evaluation through the Tchebychev inequality that looses also part of the tightness. However, as opposed to all the other methods described above, this method cannot claim theoretical correctness of its results.

In the light of similarity of the proposed scheme to the pure empirical test, we can make a direct comparison of the computational cost of the two tests. See the details in the Section 5.4.

5. Experimental Results

5.1 Test-Cases to Study

We carry out a number of tests on each of the three following dictionaries:

1. $\mathbf{D}$ – *Random* is the dictionary of size 128×256, which consists of ℓ_2-normalized random vectors, independently drawn from the Normal distribution on the unit sphere. Such a dictionary is often used in numerical experiments as well as in various appli-

cations.

2. **D** − *Spoiled* is the dictionary **D** − *Random*, which has undergone an operation designed to create a small set of columns with high linear dependence. More precisely, we re-generate a set of 3 columns as a random linear combination of 12 other columns. This dictionary is used to demonstrate the ability of the *capacity-sets* methods to better handle dictionaries with a non-uniform distribution of inner products.
3. **D** − *DCT* is the orthonormal pair $[\mathbf{I}, \mathbf{C}^*]$ of size 128×256, where $\mathbf{C}$ is the 1-dimensional Discrete Cosine basis and $\mathbf{I}$ the identity matrix.

5.2 Behavior of q and Q

As explained earlier, the passage from the capacity vector $\mathbf{q}$ to the matrix $\mathbf{Q}$ was motivated by the fact that $Q_{i,j}$ provide a lower bound in this context. To exhibit the numerical behavior of these bounds, we compute the mean and the variance of the family of ratios

$$R_{k,l} = \frac{Q_{k,l}}{q_k + q_l} \quad \text{for} \quad k \neq l \in \Omega. \tag{5.1}$$

The mean and variance of these ratios for the three test cases is given in Table 1.1.

As these figures show, we earn up to 30% of the upper bound value by upgrading to Capacity Matrix from the Capacity Vector. This ratio between the two bounds for the corresponding indices is very stable, as seen from the low values of the standard deviation $\sigma(R)$.

To display the power of Conjecture B, we show that $E_Q < 2E_q$ and either $\sigma_Q^2 < 2\sigma_q^2$ or $\sigma_Q^2 \ll E_Q^2$. The corresponding values for various dictionaries are presented in the table below.

Dictionary	$\mathbb{E}(R)$	$\sigma(R)$
$\mathbf{D}-Random$	0.7175	0.0008
$\mathbf{D}-Spoiled$	0.7154	0.001
$\mathbf{D}-DCT$	0.6509	0.0109

Table 1.1

Behavior of the *capacity-sets* **q** and **Q** by evaluating the mean and variance of the ratios.

Dictionary	E_Q	$2E_q$	σ_Q^2	$2\sigma_q^2$
D-Random 32×128	0.2329	0.3179	0.5849e-3	0.8252e-3
D-Random 64×128	0.1695	0.2345	0.1405e-3	0.1654e-3
D-Random 128×256	0.1235	0.1721	0.4511e-4	0.5652e-4
D-DCT 64×128	0.1687	0.2586	0.4732e-3	0.0112e-3
D-DCT 128×256	0.1265	0.1943	0.4070e-3	0.4144e-5

Table 1.2

Comparison of mean and variance of capacity sets.

Notice that for the $\mathbf{D}-DCT$ dictionary the variance of the capacity vector is smaller than that of the Capacity matrix, due to the special structure of this dictionary. Nevertheless, as seen later in the results section, Conjecture B predicts BP success on support sizes larger than those allowed by Theorem A.

5.3 Compared Methods

We perform a number of computations, applying various methods for the estimation of BP performance on the given dictionaries. The results are expressed via a set of Estimation Functions, $EF : \Omega \to \mathbb{R}$, which value at $\ell \in \Omega$ is the predicted percentage of ℓ-sized supports which admit recovery by ℓ_1-norm optimization. The EFs considered are the following:

1. EF-emp - The standard empirical test on the dictionary. This test is done by drawing $1,000$ random supports for each cardinality ℓ, generating a corresponding signal, and solving the BP per each. EF-emp is obtained by showing the relative number of successes in recovering the support.
2. EF-CB - the classical coherence-based upper bound $\frac{1}{2}(1+\frac{1}{\mu})$, provided by the Theorem 2.2.
3. EF-thmA - expresses the results of the Theorem A, EF-thmA $(\ell) = P(\ell)$ as defined in the statement of the theorem. The values are computed from $\mathbf{q}$ of the dictionary.
4. EF-thmB - expresses the results of the Conjecture B, computed from the capacity matrix $\mathbf{Q}$ of the dictionary.
5. EF-compB - The results of the sampling algorithm based on $\mathbf{Q}$, which results support the estimation of Conjecture B (see section 4.2).
6. EF-GB - The Grassmanian upper bound, computed by the formula for the Classical Bound using the ideal coherence $\mu = \sqrt{\frac{L-N}{N(L-1)}}$.

This last EF deserves more explanation: Among all possible dictionaries of size $N \times L$, the Grasssmanian frame is the one leading to the smallest possible coherence $\mu = \sqrt{\frac{L-N}{N(L-1)}}$ [17]. Thus, this leads to the most optimistic worst-case bound. When the dictionary is “un-balanced”, implying a large spread of inner-products in the Gram-matrix, we know that the *mutual-coherence*-bound deteriorates dramatically. Thus, by using the Grassmanian Bound, we test what is the best achievable coherence-based performance behavior for the same dictionary size.

5.4 Complexity Analysis of the Methods

We argue the usefulness of Capacity-based numerical algorithms for an evaluation of a given dictionary $\mathbf{D}$. To that end, we consider the computational complexity of each method listed in previous section.

1. EF-emp - The standard empirical test of $\mathbf{D}$ is conveyed as follows: for each support size ℓ, pick $M >> L$ random subsets Γ of columns of size ℓ. For each Γ, generate a signal with random coefficients vector supported on Γ and test if BP will recover the support. Since in practice maximal relevant size ℓ is proportional to L, the computational complexity of this test is $\mathcal{O}(M \cdot L \cdot C_{LP}(L))$, where $C_{LP}(L)$ denotes the complexity of linear programming algorithm for problem of size L.
2. EF-CB requires the computation of μ, which takes $\mathcal{O}(L \cdot N)$ flops.
3. EF-thmA - To employ results of the Theorem A, the capacity vector $\mathbf{q}$ is computed in ($\mathcal{O}(L \cdot C_{LP}(L))$), and then for each ℓ the probability $P(\ell)$, defined in the statement of Theorem A, is computed in $\mathcal{O}(L)$. Overall complexity - $\mathcal{O}(L^2 + L \cdot C_{LP}(L)) = \mathcal{O}(L \cdot C_{LP}(L))$.
4. EF-thmB - To employ results of the Conjecture B, the capacity vector $\mathbf{q}$ is computed in ($\mathcal{O}(L^2 \cdot C_{LP}(L))$), and then for each ℓ the probability $P(\ell)$, defined in the statement of Conjecture B, is computed in $\mathcal{O}(L^2)$. Overall complexity - $\mathcal{O}(L^3 + L^2 \cdot C_{LP}(L)) = \mathcal{O}(L^2 \cdot C_{LP}(L))$.
5. EF-compB - Our heaviest (and best-performance) algorithm conducts a semi-empirical test: for each support size ℓ, pick

$M >> L$ random subsets of columns of size ℓ, and employ the analysis detailed in 4.2. The computational cost of single support treatment is $\mathcal{O}(\ell^2 \cdot log(\ell))$. Overall complexity is $\mathcal{O}(L^2 \cdot C_{LP}(L) + M \cdot L^2 \cdot log(Ł))$.

As seen from the analysis above, only the EF-compB has non-negligible computational complexity. When comparing EF-emp and EF-compB, we can concentrate on the relative complexities of linear programming solver versus the $\mathcal{O}(\ell^2 \cdot log(\ell))$ of the partition algorithm, and the benefit of the later is evident.

5.5 Comparison Results

Figure 1 presents the obtained graphs of the various EF-s functions described above, for the three dictionaries described at the top of this section. As we see from the left-side graphs in the figures, for all the dictionaries the empirically established support size which admits BP recovery is at least 40 columns. Note that this relative number of columns is also predicted in [10], however, this holds true only asymptotically (for dictionaries of growing sizes) and for specific random dictionaries.

Returning to statements which hold for our modest size of 128×256, we notice that the estimation made by the sampling algorithm based on the Capacity Matrix (EF-compB) is much better than the Classical bound, established so far in the literature. The difference is especially high for the D-Spoiled dictionary, which reflects the fact that methods based on *capacity sets* manage well the non-uniform distribution of inner products.

On the right side of each figure we display various method developed in this work. Noticeably, the results of Conjecture B(EF-thmB)

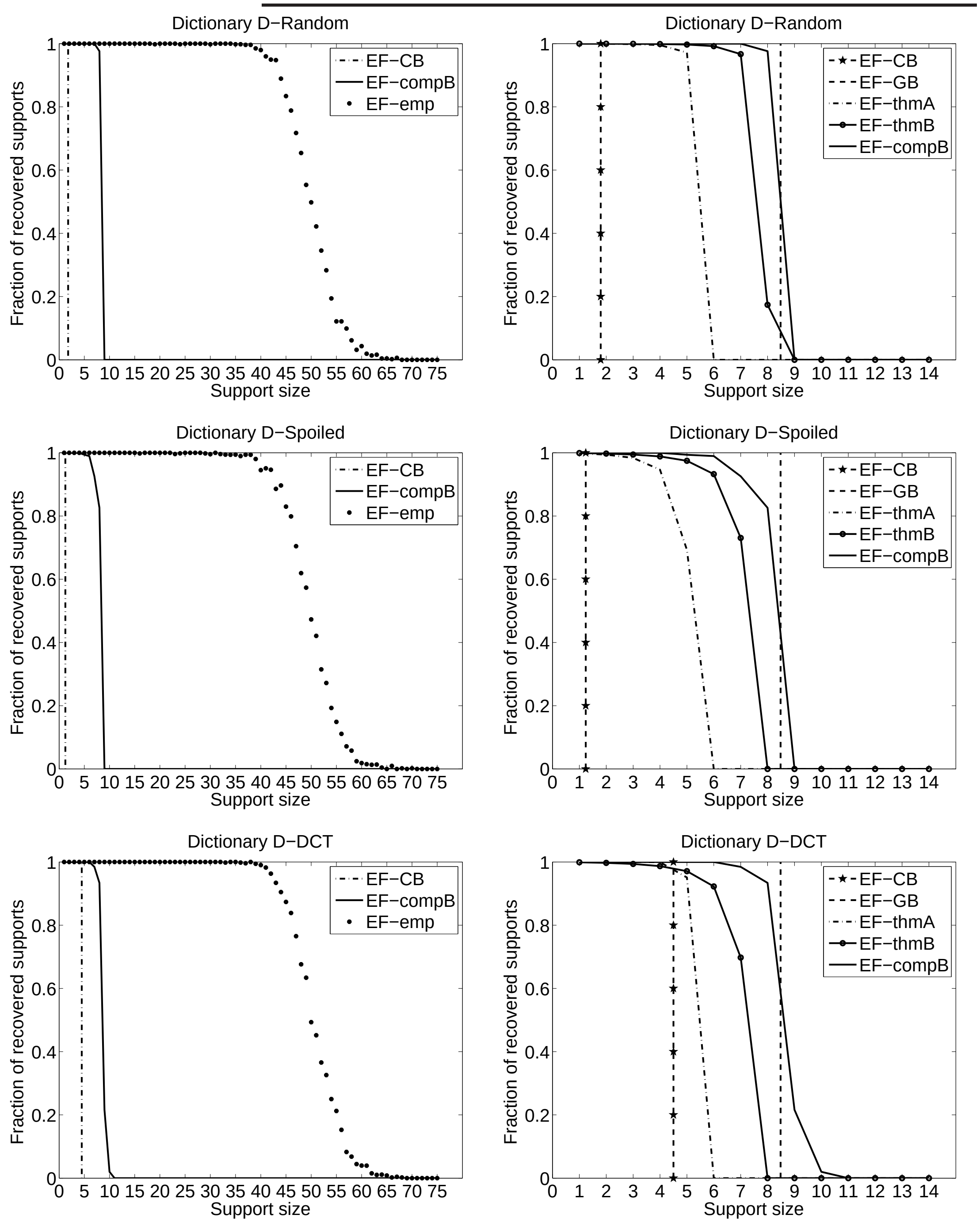

Figure 1: Estimation Functions for various dictionaries of size 128×256.

are stronger than those of Theorem A (EF-thmA), which is explained by the benefit of using the Capacity Matrix rather than the Capacity Vector. This benefit is expressed in the ratio values given in Tables 1.1, 1.2 and explained thereafter. Apparently, Conjecture B does not express the full power of the Capacity Matrix estimation, since the sampling algorithm based on its values (EF-compB) outperforms EF-thmB by $15-20\%$. This algorithm produces values which are quite close to the Grassmanian Bound, the best possible bound one can hope to obtain using coherence-based estimation for the given dictionary size. We do not have enough information to explain the fact that values of EF-compB and of Grassmanian bound nearly coincide for all the dictionaries discussed here (and additional ones examined during the work); Discovering the reason underlying this connection may be a lead to important insights regarding the Basis Pursuit performance.

Appendix A

We prove the claim 3.4.

Theorem C. *For the two random variables, x_ℓ and y_ℓ, defined in 3.3, the following relations between the first and second moments hold:*

$$\mathbb{E}(x_\ell) = \mathbb{E}(y_\ell) \quad and \quad var(x_\ell) \leq var(y_\ell). \tag{A-1}$$

Proof: We begin by introducing some notation. Fix the support size $1 \leq \ell \leq L$. For any $1 \leq k \leq \ell$, we denote by $\mathcal{C}_\ell^k$ the collection of all ℓ-sized non-ordered multisets of indices from Ω (with repetitions), which have precisely k distinct elements each. For instance, $\{1,4,5,4,7\}$ and $\{5,1,7,4,4\}$ are two distinct elements of $\mathcal{C}_5^4$. Such multiset will be sometimes referred to as "index set". Also, we define $\mathcal{D}_\ell^n = \mathcal{C}_\ell^\ell \cup \mathcal{C}_\ell^{\ell-1} \cup$

$... \cup \mathcal{C}_\ell^{\ell-n}$, the collection of all ℓ-sized multisets having at least $\ell - n$ distinct elements.

In this notation, x_ℓ is a random variable with uniform distribution over the domain $\mathcal{D}_\ell^0$, which admits value $\sum_{k\in\Lambda} q_k$ on a given element $\Lambda \in \mathcal{D}_\ell^0$. The variable y_ℓ has the same definition on a larger domain $\mathcal{D}_\ell^{\ell-1}$, containing the domain of x_ℓ. Therefore, we treat both x_ℓ and y_ℓ as restrictions of the same uniformly distributed random variable x on the corresponding domains: $x_\ell = x_{|\mathcal{D}_\ell^0}$, $y_\ell = x_{|\mathcal{D}_\ell^{\ell-1}}$. In the proof we use the following basic property of the variance:

Proposition 5.1. *Let z be a random variable defined over a domain given as the disjoint union $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup ... \cup \mathcal{D}_n$, with uniform distribution. Denote $v = var(z_{|\mathcal{D}}), v_i = var(z_{|\mathcal{D}_i}), s_i = |\mathcal{D}_i|$. Then $v = \dfrac{\sum_{i=1}^n s_i v_i}{\sum_{i=1}^n s_i}$.*

Part 1. The expectation of the random variable x restricted to $\mathcal{D}_\ell^0$ is computed by

$$\mathbb{E}(x_{|\mathcal{D}_\ell^0}) = \frac{1}{|\mathcal{D}_\ell^0|} \sum_{\Lambda\in\mathcal{D}_\ell^0} \sum_{k\in\Lambda} q_k.$$

This sum contains $|\mathcal{D}_\ell^0| \cdot \ell$ elements, and for each $j \in \Omega$, q_j appears in it the same number of times. Therefore, each q_j appears $|\mathcal{D}_\ell^0| \frac{\ell}{L}$ times, and we have $\mathbb{E}(x_{|\mathcal{D}_\ell^0}) = \dfrac{\ell}{L} \displaystyle\sum_{k\in\Omega} q_k = \ell E_q$. The mean of $x_{|\mathcal{D}_\ell^{\ell-1}}$ is computed similarly:

$$\mathbb{E}(x_{|\mathcal{D}_\ell^{\ell-1}}) = \frac{1}{|\mathcal{D}_\ell^{\ell-1}|} \sum_{\Lambda\in\mathcal{D}_\ell^{\ell-1}} \sum_{k\in\Lambda} q_k.$$

Here each q_j appears $|\mathcal{D}_\ell^{\ell-1}| \frac{\ell}{L}$ times, and we have $\mathbb{E}(x_{|\mathcal{D}_\ell^{\ell-1}}) = \dfrac{\ell}{L} \displaystyle\sum_{k\in\Omega} q_k = \ell E_q$.

This proves our first claim, $\mathbb{E}(x_\ell) = \mathbb{E}(y_\ell)$. For the rest of the proof, where only the variance of the two variables is considered, we

assume w.l.g. that the expectation of x_ℓ and y_ℓ is zero (in the light of equality $var(z) = var(z - \mathbb{E}(z)$ for any random variable z), that is $E_q = 0$.

Part 2. We consider the extension of x, defined so far on domain comprising of distinct ℓ-sized index sets, to the domain where each such set may appear any finite number of times. x still has a uniform distribution over this collection. Thus, a disjoint union of two or more (non-necessarily distinct) index sets is a sub-domain to which x may be restricted.

For any $0 \leq n < \ell$, we define two disjoint unions

$$\mathcal{A}_n = \bigcup_{\Gamma \in \mathcal{D}^n_{\ell-1}} \{\Gamma \cup \{j\} \mid j \in \Gamma\},$$
$$\mathcal{B}_n = \bigcup_{\Gamma \in \mathcal{D}^n_{\ell-1}} \{\Gamma \cup \{j\} \mid j \in \Omega\}$$

(In the definition of $\mathcal{A}_n$, the set $\Gamma \cup \{j\}$ is added to the collection one time for each appearance of j in Γ.)

Let $\Lambda \in \mathcal{C}^k_\ell$ be a set which contains distinct indices $j_1, ..., j_k$ with multiplicities $m_1, .., m_k$ (so that $\sum_{i=1}^k m_i = \ell$). For each $1 \leq i \leq k$, Λ is obtained in $\mathcal{A}_n$ $m_i - 1$ times in the form $\Gamma \cup \{j_i\}$ for an appropriate $\Gamma = \Gamma_i \in \mathcal{C}^k_{\ell-1}$ (this claim also holds vacuously for $m_i = 1$). Therefore, the number of copies of Λ in $\mathcal{A}_n$ equals $\sum_{i=1}^k (m_i - 1) = \ell - k$. Also, Λ appears in $\mathcal{B}_n$ precisely once for each $j_1, ..., j_k$, in the form $\Gamma \cup \{j_i\}$ (for an appropriate $\Gamma = \Gamma_i$ each time). Therefore, $\mathcal{B}_n$ contains k copies of Λ.

Denote a disjoint union of a distinct copies of some collection $\mathcal{C}$ by $a \cdot \mathcal{C}$. Then we can write $\mathcal{A}_n, \mathcal{B}_n$ as

$$\mathcal{A}_n = 0 \cdot \mathcal{C}^\ell_\ell \cup 1 \cdot \mathcal{C}^{\ell-1}_\ell \cup ... \cup n \cdot \mathcal{C}^{\ell-n}_\ell \tag{A-2}$$

$$\mathcal{B}_n = \ell \cdot \mathcal{C}_\ell^\ell \cup (\ell-1) \cdot \mathcal{C}_\ell^{\ell-1} \cup ... \cup (\ell-n) \cdot \mathcal{C}_\ell^{\ell-n} \tag{A-3}$$

We prove the following inequality:

$$var(x_{|\mathcal{B}_n}) \leq var(x_{|\mathcal{A}_n}).$$

Since $E_q = 0$ by our assumption, the expectations of $x_{|\mathcal{A}_n}$ and $x_{|\mathcal{B}_n}$ also equal zero: by the argument similar to one presented in the first part of the proof, $\mathbb{E}(x_{|\mathcal{A}_n}) = \mathbb{E}(x_{|\mathcal{B}_n}) = \ell \cdot E_q$. Thus we have

$$var(x_{|\mathcal{A}_n}) = \frac{1}{|\mathcal{D}_{\ell-1}^n|} \sum_{\Gamma \in \mathcal{D}_{\ell-1}^n} \frac{1}{\ell-1} \sum_{j\in\Gamma} (\sum_{k\in\Gamma} q_k + q_j)^2.$$

For the brevity of the argument we introduce the notation $q_\Gamma = \sum_{k\in\Gamma} q_k$. Then $var(x_{|\mathcal{A}_n})$ reads as

$$\begin{aligned} var(x_{|\mathcal{A}_n}) &= \frac{1}{|\mathcal{D}_{\ell-1}^n|} \sum_{\Gamma \in \mathcal{D}_{\ell-1}^n} \frac{1}{\ell-1} \sum_{j\in\Gamma} (q_\Gamma^2 + q_j^2 + 2q_\Gamma q_j) = \\ &= \frac{1}{|\mathcal{D}_{\ell-1}^n|} \sum_{\Gamma \in \mathcal{D}_{\ell-1}^n} q_\Gamma^2 + \frac{1}{\ell-1} \sum_{j\in\Gamma} (q_j^2 + 2q_\Gamma q_j). \end{aligned}$$

Similarly, we have

$$\begin{aligned} var(x_{|\mathcal{B}_n}) &= \frac{1}{|\mathcal{D}_{\ell-1}^n|} \sum_{\Gamma \in \mathcal{D}_{\ell-1}^n} \frac{1}{L} \sum_{j\in\Omega} (\sum_{k\in\Gamma} q_k + q_j)^2 = \\ &= \frac{1}{|\mathcal{D}_{\ell-1}^n|} \sum_{\Gamma \in \mathcal{D}_{\ell-1}^n} q_\Gamma^2 + \frac{1}{L} \sum_{j\in\Omega} (q_j^2 + 2q_\Gamma q_j). \end{aligned}$$

The summand $\frac{1}{|\mathcal{D}_{\ell-1}^n|} \sum_{\Gamma \in \mathcal{D}_{\ell-1}^n} q_\Gamma^2$ appears in both expressions hence cancels out. We consider the term $\frac{1}{|\mathcal{D}_{\ell-1}^n|} \sum_{\Gamma \in \mathcal{D}_{\ell-1}^n} \frac{1}{\ell-1} \sum_{j\in\Gamma} q_j^2$ in $var(x_{|\mathcal{A}_n})$. The element q_a^2 appears in it same number of times for every $a \in \Omega$. Hence $\frac{1}{|\mathcal{D}_{\ell-1}^n|} \sum_{\Gamma \in \mathcal{D}_{\ell-1}^n} \frac{1}{\ell-1} \sum_{j\in\Gamma} q_j^2 = \frac{1}{L} \sum_{a\in\Omega} q_a^2$. By same argument, in

the expression of $var(x_{|\mathcal{B}_n})$ we have $\frac{1}{|\mathcal{D}^n_{\ell-1}|}\sum_{\Gamma\in\mathcal{D}^n_{\ell-1}}\frac{1}{L}\sum_{j\in\Omega}q_j^2 = \frac{1}{L}\sum_{a\in\Omega}q_a^2$, hence this quadratic term also cancels out. In the light of these observations, we obtain

$$var(x_{|\mathcal{A}_n}) - var(x_{|\mathcal{B}_n}) = \frac{2}{|\mathcal{D}^n_{\ell-1}|}\sum_{\Gamma\in\mathcal{D}^n_{\ell-1}} q_\Gamma(\frac{1}{\ell-1}\sum_{i\in\Gamma}q_i - \frac{1}{L}\sum_{j\in\Omega}q_j).$$

Here we substitute again q_Γ for $\sum_{i\in\Gamma}q_i$ and recall $\frac{1}{L}\sum_{j\in\Omega}q_j = E_q = 0$. Thus, we have

$$var(x_{|\mathcal{A}_n}) - var(x_{|\mathcal{B}_n}) = \frac{2}{(\ell-1)|\mathcal{D}^n_{\ell-1}|}\sum_{\Gamma\in\mathcal{D}^n_{\ell-1}} q_\Gamma^2 \geq 0.$$

In order to use this result for the proof of the theorem, we make the following observations : Denote $v_n = var(x_{|\mathcal{C}^n_\ell})$ and $s_n = |\mathcal{C}^n_\ell|$. By virtue of the decomposition (A-2), $var(x_{|\mathcal{A}_n})$ can be written as $var(x_{|\mathcal{A}_n}) = \frac{\sum_{i=0}^n i\cdot s_{\ell-i}v_{\ell-i}}{\sum_{i=0}^n i\cdot s_{\ell-i}}$ (see Proposition 5.1). Similarly, we have $var(x_{|\mathcal{B}_n}) = \frac{\sum_{i=0}^n(\ell-i)\cdot s_{\ell-i}v_{\ell-i}}{\sum_{i=0}^n(\ell-i)\cdot s_{\ell-i}}$. We compute the coefficients of v_i in the expression

$$var(x_{|\mathcal{A}_n}) - var(x_{|\mathcal{B}_n}) = \frac{\sum_{i=0}^n i\cdot s_{\ell-i}v_{\ell-i}}{\sum_{i=1}^n i\cdot s_{\ell-i}} - \frac{\sum_{i=0}^n(\ell-i)\cdot s_{\ell-i}v_{\ell-i}}{\sum_{i=1}^n(\ell-i)\cdot s_{\ell-i}}.$$

For any $0\leq k\leq n$, the coefficient of v_{l-k} is

$$\frac{1}{Den}s_{\ell-k}\left(k\sum_{i=1}^n(\ell-i)\cdot s_{\ell-i} - (\ell-k)\sum_{i=1}^n i\cdot s_{\ell-i}\right)$$
$$= \frac{1}{Den}\ell\cdot s_{\ell-k}\sum_{i=0}^n(k-i)s_{\ell-i},$$

with

$$Den = \sum_{i=1}^n i\cdot s_{\ell-i}\cdot\sum_{i=1}^n(\ell-i)\cdot s_{\ell-i}.$$

We denote $\alpha_{\ell-k} = \ell \sum_{i=0}^{n} (k-i) s_{\ell-i}$, for $1 \leq k \leq n$, in order to write the above difference as

$$0 \leq var(x_{|\mathcal{A}_n}) - var(x_{|\mathcal{B}_n}) = \frac{1}{Den} \sum_{k=0}^{n} \alpha_{\ell-k} s_{\ell-k} v_{\ell-k}. \tag{A-4}$$

The constant $\frac{1}{Den}$ is positive, since $n < \ell$. Thus,it can be omitted while preserving the inequality:

$$0 \leq \sum_{k=0}^{n} \alpha_{\ell-k} s_{\ell-k} v_{\ell-k}. \tag{A-5}$$

The coefficients in this expression have the two following properties:

1. $\sum_{k=0}^{n} s_{\ell-k} \alpha_{\ell-k} = 0$.

2. $\forall j, \alpha_{j-1} - \alpha_j = \ell \sum_{i=0}^{n} s_{\ell-i}$.

To show the first equality, we consider the sum in (1) as the linear combination of the elements $s_{\ell-i} s_{\ell-j}$, $i, j = 0, ..., n$. The coefficient of $s_{\ell-i} s_{\ell-i}$ is zero for any i. For any $i \neq j$, $s_{\ell-i} s_{\ell-j}$ appears just in two components of the sum above, namely, $s_{\ell-i} \alpha_{\ell-i}$ and $s_{\ell-j} \alpha_{\ell-j}$. Specifically, $\alpha_{\ell-i}$ contains the summand $\ell(i-j) s_{\ell-j}$, and $\alpha_{\ell-j}$ contains the summand $\ell(j-i) s_{\ell-i}$, therefore in the sum $s_{\ell-i}\alpha_{\ell-i} + s_{\ell-j}\alpha_{\ell-j}$ the coefficient of $s_{\ell-i} s_{\ell-j}$ is zero. The second property follows from the definition of α_i. In the light of the first property, A-5 can be written as

$$(\sum_{k=1}^{n} \alpha_{\ell-k} s_{\ell-k}) v_\ell \leq \sum_{k=1}^{n} \alpha_{\ell-k} s_{\ell-k} v_{\ell-k}. \tag{A-6}$$

Equipped with these observations, we prove, by induction on n, the inequality

$$var(x_{|\mathcal{D}_\ell^0}) \leq var(x_{|\mathcal{D}_\ell^n}).$$

for any $n = 1, ..., \ell - 1$. The theorem follows for $n = \ell - 1$. By Proposition 5.1, $var(x_{|\mathcal{D}_\ell^n}) = \frac{\sum_{i=0}^n s_{\ell-i} v_{l-i}}{\sum_{i=0}^n s_{\ell-i}}$, and $var(x_{|\mathcal{D}_\ell^0})$ is just v_ℓ. Thus we need to prove

$$v_\ell \leq \frac{\sum_{i=0}^n s_{\ell-i} v_{l-i}}{\sum_{i=0}^n s_{\ell-i}},$$

or

$$(\sum_{i=1}^n s_{\ell-i}) v_\ell \leq \sum_{i=1}^n s_{\ell-i} v_{l-i}. \tag{A-7}$$

For $n = 1$, A-6 reads as

$$\alpha_{\ell-1} s_{\ell-1} v_\ell \leq \alpha_{\ell-1} s_{\ell-1} v_{\ell-1}.$$

Here $\alpha_{\ell-1} = \ell s_\ell > 0$, thus we obtain the inequality

$$s_{\ell-1} v_\ell \leq s_{\ell-1} v_{\ell-1},$$

as required. Now, we assume by induction that inequality A-7 holds up to $n - 1$ and prove for n. We use (A-6):

$$(E1): \quad (\sum_{k=1}^n \alpha_{\ell-k} s_{\ell-k}) v_\ell \leq \sum_{k=1}^n \alpha_{\ell-k} s_{\ell-k} v_{\ell-k}.$$

This inequality undergoes a series of transformations designed to bring it to the form of A-7.

First, we have $\alpha_{\ell-1} < \alpha_{\ell-2}$. Since $v_\ell \leq v_{\ell-1}$ by the proof for $n = 1$, we have an inequality

$$(d1): \quad (\alpha_{\ell-2} - \alpha_{\ell-1}) s_{\ell-1} v_\ell \leq (\alpha_{\ell-2} - \alpha_{\ell-1}) s_{\ell-1} v_{\ell-1}$$

Adding $(d1)$ to the inequality $(E1)$, we arrive at

$$(E2): \quad \left(\alpha_{\ell-2}(s_{\ell-1} + s_{\ell-2}) + \sum_{k=3}^n \alpha_{\ell-k} s_{\ell-k} \right) v_\ell \leq$$
$$\leq \alpha_{\ell-2}(s_{\ell-1} v_{\ell-1} + s_{\ell-2} v_{\ell-2}) + \sum_{k=3}^n \alpha_{\ell-k} s_{\ell-k} v_{\ell-k}.$$

Second, by induction assumption for $n = 2$ we have the inequality

$$(s_{\ell-1} + s_{\ell-2})v_\ell \leq s_{\ell-1}v_{\ell-1} + s_{\ell-2}v_{\ell-2}.$$

Also, $\alpha_{\ell-2} \leq \alpha_{\ell-3}$ as noticed earlier. Then we can construct the next inequality in order to add it to $(E2)$:

$$(d1): \quad (\alpha_{\ell-3}-\alpha_{\ell-2})(s_{\ell-1}+s_{\ell-2})v_\ell \leq (\alpha_{\ell-3}-\alpha_{\ell-2})(s_{\ell-1}v_{\ell-1}+s_{\ell-2}v_{\ell-2})$$

This results in the following expression:

$$(E3): \qquad \left(\alpha_{\ell-3}\sum_{i=1}^{3} s_{\ell-i} + \sum_{k=4}^{n} \alpha_{\ell-k}s_{\ell-k}\right)v_\ell \leq$$
$$\leq \alpha_{\ell-3}\sum_{i=1}^{3}(s_{\ell-i}v_{\ell-i}) + \sum_{k=4}^{n} \alpha_{\ell-k}s_{\ell-k}v_{\ell-k}.$$

In this fashion we make $n-1$ steps resulting in the inequality

$$(E(n)): \quad (\alpha_{\ell-n}\sum_{i=1}^{n} s_{\ell-i})v_\ell \leq \alpha_{\ell-n}\sum_{i=1}^{n} s_{\ell-i}v_{\ell-i}$$

Notice that $\alpha_{\ell-n}$ is positive: $\alpha_{\ell-n} = s_{\ell-n}\ell(ns_\ell+(n-1)s_{\ell-1}+...+s_{\ell-n+1})$. Thus, we obtain the desired result. As mentioned, the theorem follows for $n = \ell - 1$. $\square$

Appendix B

We prove the equality of expectations

$$\mathbb{E}(x_\ell) = \mathbb{E}(y_\ell), \tag{B-1}$$

for random variables x_ℓ and y_ℓ defined in the proof of Conjecture B. Recall that y_ℓ is a sum of $\frac{\ell}{2}$ values from $\mathbf{Q}$, uniformly distributed over this matrix, therefore $\mathbb{E}(y_\ell) = \frac{\ell}{2}\mathbb{E}_Q$. We show $\mathbb{E}(x_\ell) = \frac{\ell}{2}\mathbb{E}_Q$, too, by considerations of symmetry, similar to those used in the proof of Theorem A, part 1.

Namely, we consider a totality $\mathcal{P}_\ell$ of partitions of all ℓ-sized supports $\Lambda \subset \Omega$, into ordered pairs of indices. An element in this collection is therefore a pair $(\Lambda, \mathcal{I}_\Lambda)$. We clarify that the index sets $\Lambda \subset \Omega$ are chosen without repetitions and up to a permutation of their elements. Now, let (i, j) be an ordered pair of indices from Ω. We argue that the number of appearances of this pair in the elements of $\mathcal{P}_\ell$ does not depend on choice of i and j. Indeed, this number is just the size of the collection $\mathcal{P}_{\ell-2}$, built for submatrix of $\mathbf{Q}$ with i-th and j-th rows and columns missing.

Since $x_\ell(\Lambda, \mathcal{I}_\Lambda)$ is the sum $\sum_{(i,j)\in\mathcal{I}_\Lambda} \mathbf{Q}(i, j)$, we conclude that all the elements $\mathbf{Q}(i, j)$ contribute to the value of x_ℓ with equal probability, hence $\mathbb{E}(x_\ell) = \frac{\ell}{2}\mathbb{E}_Q$ as desired.

Now we provide an empirical evidence to the claim

$$var(x_\ell) \leq var(y_\ell) \tag{B-2}$$

Statistical data below supports this inequality. While the variance of y_ℓ is known precisely, for x_ℓ we estimate it by drawing 10^4 random subsets of indices for each support size up to half the signal dimension of the dictionary. Results are presented in Figure 2. The computation is carried out for a number of dictionary sizes on dictionary $\mathbf{D}$-Random. As can be seen from these figures, the gap between $var(x_\ell)$ and $var(y_\ell)$ is roughly proportional to the support size.

Same experiments on dictionary $\mathbf{D}$-DCT display different results: the variance of both variables coincides. As number of samples grows, we observe that the difference of variance values, for all support sizes, tends to zero. We conclude that for this specific dictionary, B-2 is an equality.

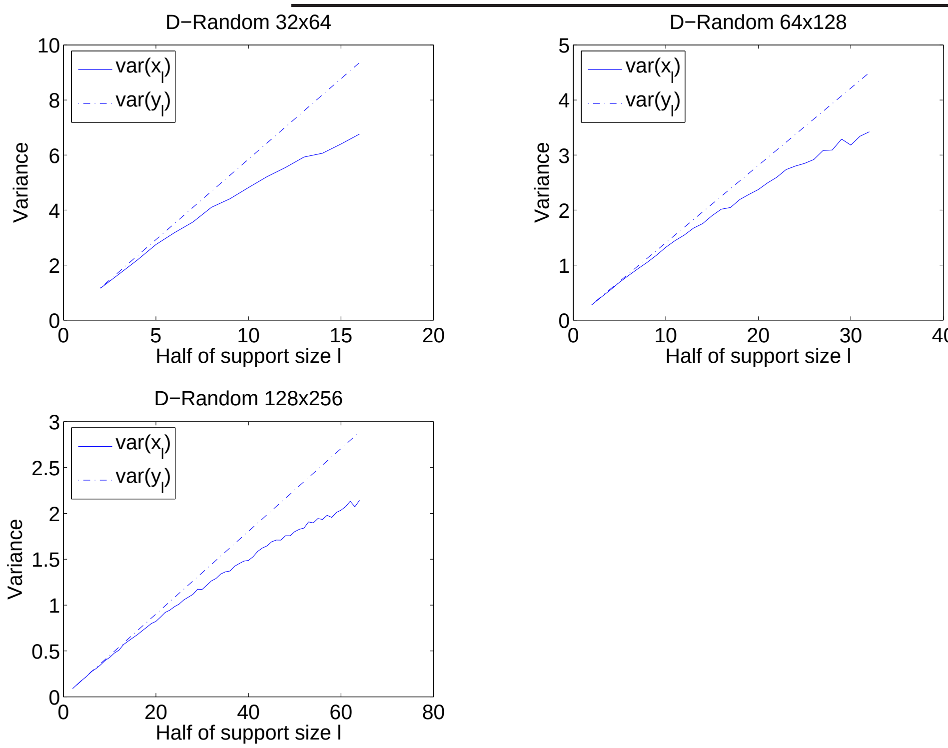

Figure 2: The variances of x_ℓ and y_ℓ (scaled by 10^3)

References

[1] E.J. Candès, J. Romberg, and T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. on Information Theory*, vol. 52(2), pp. 489–509, 2006.

[2] E. Candès, J. Romberg, Quantitative robust uncertainty principles and optimally sparse decompositions, *Foundations of Computational Mathematics*, Vol. 6, Issue 2, April 2006.

[3] E.J. Candès and T. Tao, Decoding by linear programming, *IEEE Trans. Inform. Theory*, vol. 51, pp. 4203–4215, 2005.

[4] S.S. Chen, Basis pursuit, Ph.D. dissertation, Stanford Univ., Stanford, CA, Nov. 1995

[5] S.S. Chen, D. L. Donoho, and M. A. Saunders, Atomic decomposition by basis pursuit, *SIAM J. Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1999.

[6] D.L. Donoho, For most large underdetermined systems of linear equations the minimal l^1-norm solution is also the sparsest solution, *Communications On Pure And Applied Mathematics*, vol. 59, No. 6, pp. 797–829, June 2006.

[7] D.L. Donoho, Neighborly polytopes and sparse solution of underdetermined linear equations, Technical report. Stanford University, Department of Statistics, #2005-04 (2005)

[8] D.L. Donoho and M. Elad, Optimally sparse representation in general (non- orthogonal) dictionaries via l^1 minimization,*Proceedings of the National Academy of Sciences*, vol 100(5), pp. 2197–2202, 2003.

[9] D.L. Donoho and X. Huo, Uncertainty principles and ideal atomic decomposition, *IEEE Trans. On Information Theory*, 47(7):2845–2862, 1999.

[10] D.L. Donoho and J. Tanner, Thresholds for the recovery of sparse solutions vial L1 Minimization, *40th Annual Conference on Information Sciences and Systems,* 2006

[11] M. Elad and A.M. Bruckstein, A generalized uncertainty principle and sparse representations in pairs of bases, *IEEE Trans. Info. Thry*, vol. 49, pp. 2558–2567, 2002

[12] J.J. Fuchs, On sparse representations in arbitrary redundant bases, *IEEE Trans. on Information Theory*, vol. 50(6), pp. 1341–1344, 2004.

[13] R. Gribonval and M. Nielsen, Sparse decompositions in unions of bases, *IEEE Trans. on Information Theory*, vol 49(12), pp. 3320–3325, 2003.

[14] G.H. Hardy, J.E. Littlewood, and G. Pólya, *Inequalities*, 2nd ed. Cambridge, England: Cambridge University Press, pp. 43–45 and 123, 1988.

[15] D.M. Malioutov, M. Cetin, and A.S. Willsky, Optimal sparse representations in general overcomplete bases, *IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP*, May 2004, Montreal, Canada.

[16] B.K. Natarajan, Sparse approximate solutions to linear systems, *SIAM Journal on Computing*, vol. 24, pp. 227–234, 1995.

[17] T. Strohmer and R.W. Heath, Jr., Grassmannian Frames with applications to coding and communications, *Applied and Computational Harmonic Analysis*, vol. 14, Issue 3, pp. 257–275, May 2003.

[18] T. Tao, An uncertainty principle for cyclic groups of prime order, *Math. Res. Lett.*, vol. 12, no. 1, pp 121–127, 2005.

[19] J. Tropp, Random subdictionaries of random dictionaries, Preprint, 2006.

[20] J. Tropp, Greed is good: algoritmic results for sparse approximation, *IEEE Trans. Inform. Theory*, vol. 50, no. 10, pp. 2231–2242, October 2004.